DOCUMENT RESUME

ED 448 211 TM 032 241

AUTHOR	McDonald, Jo-Anne; Hall, Lisa
TITLE	The Impact of Instructional Set on Distributions of Self-Report Ratings on a Survey of Personality Characteristics.
PUB DATE	2000-04-00
NOTE	31p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Employees; Item Response Theory; *Multivariate Analysis; *Personality Measures; Rating Scales; Responses; *Selection
IDENTIFIERS	*Instructions; *Item Parameters; Rasch Model; Self Report Measures

ABSTRACT

The purpose of this study was to examine the effect of instrument completion instructions on univariate and multivariate distributional characteristics and relationships among variables. Instructions allowed free-choice allotment of ratings (unforced instructions) or requested the subject to assign a certain number of ratings to either the highest or lowest rating-scale categories (forced instructions). A total of 126 participants using unforced distributions completed a self-report survey of personality characteristics in a utility company. The comparison sample of 346 forced distributions was collected from 5 companies. Samples of equivalent size (n=115) that contained complete data for the scale were selected for comparison. It was hypothesized that there would be effects of instructions on the central tendency and dispersion of the ratings. Although the collapsed means did not differ across instructional sets, the variability in the ratings was higher with forced-choice distributions. Standard deviations and standard errors of the mean were higher for every item answered under the forced conditions. Support for the second hypothesis, that there would be effects of instructions on the shapes of the distributions of the ratings, was found in this study. The shapes of the distributions were quite different across the sets, and apparently neither set of instructions elicited normally distributed data. The hypothesis that there would be an effect of instructions on the internal consistency of the scale was also supported by the results of the study. Five appendixes contain the instructions and frequency histograms and probability plots for some survey items. (Contains 5 tables and 18 references.) (SLD)

The Impact of Instructional Set on Distributions of Self-Report Ratings on a Survey of Personality Characteristics.

Paper presented at AERA
New Orleans, 2000

Jo-Anne McDonald
and Lisa Hall
University of Denver
Denver, Colorado

The impact of instructional set on distributions of self-report ratings on a survey of personality characteristics.

Introduction

"A statement about an empirical system is meaningful only when it is scale independent, that is, only when it is true on all of the permissible numeric scales".

-Townsend and Ashby (1984, pp. 399)

Researchers seeking to evaluate the impact of program variables on performance criteria often use survey and self-report data to do so. But often neglected is an examination of evidence supporting or conflicting with the assumptions underlying our theories and their measurement.

Statistical tests frequently assume (as do statisticians) that the level of a psychological variable is normally distributed in the populations from which samples are drawn. While most of the limited evidence points to the relatively slight impact of the violation of the normality assumption on power and significance tests (Lix, Keselman, & Keselman, 1996) as Townsend and Ashby (1984) point out, further studies are needed before we can assume these findings generalize to all applications. Indeed, researchers have found differences in results of a variety of inferential statistics in studies involving manipulations of distribution parameters, response formats, and instruction sets (e.g., Hancock & Klockars, 1996; Hough, Eaton, Dunnette, Kamp, & McCloy, 1990; Keselman & Lix, 1995). However, many of these studies used simulated data sets that conformed to a regular, known distribution.

Micceri (1989) found that, in a meta-analysis of distributional qualities in behavioral science data sets, none of over 400 data sets that he assessed passed his criteria for normality. The distribution of scores on psychometric measures, such as ability or attitude scales and achievement measures, were universally asymmetrical and "lumpy". In addition, Micceri suggested that for most data sets, non-parametric analysis might provide more accurate information and more powerful tests in social research applications.

The lack of normality in general response patterns may be compounded by social desirability and faking response styles on self-report inventories of personality or surveys of attitudes. One component of social desirability has been termed "impression management". Paulhaus (1984, 1986) believed self-report to be one way that people use to control what others think of them. They estimate situational demands and modify their behavior to enhance the probability of a preferred outcome. This parallels earlier work in social desirability by Edwards (1953). He found a high correlation (r = .87) between increase in an item's social desirability and the tendency for people to rate themselves higher on the item.

In a review of the social desirability, response bias, and faking literature, Furnham (1986) concluded that personality and adjustment measures are affected to a greater extent than other types of questionnaires. More recent reports indicate that people do fake responses and that self-report measures are particularly susceptible to faking "good" as well as faking "bad" response styles or the tendency to present one's self in a beneficial light (e.g., Barrick & Mount, 1996; Griffith, 1997; Hough, et al., 1990). However, other researchers have found that social desirability did not decrease the criterion-related validity of personality predictors of job performance (e.g., Hough, et al., 1990; Ones, Viswesvaran, & Reiss, 1996).

Another challenge for survey researchers is to assess the impact of response formats on results. Rating scales appear to be particularly susceptible to response sets more so than distally anchored or purely numeric response scales (e.g., Bardo & Yeager, 1982; Cronbach, 1950). This condition appears to hold across scale formats and number of rating categories.

Distributions of variables of interest to social scientists are often restricted, skewed, and leptokurtotic, in other words, not normally distributed. What can be done about it and how do we avoid the problem?

One solution researchers have devised is to mathematically transform the data. However, the "best" solution to the problem may be elusive. In addition, interpretations of the results are complicated by the altered state of the variables (e.g., Judd, et al., 1995; Townsend & Ashby, 1984). A second solution is to provide the opportunity for respondents to select between two statements of equal desirability. A third solution is to make the intention of the item more obscure. obscurity makes it harder for people to intentionally distort their responses. A fourth solution is to warn the respondents that their answers will be evaluated for accurateness using outside criteria. A fifth method is to include scales that detect response patterns associated with dissimulation.

Yet another approach is to circumvent the problem by manipulating the instructions for completion of an instrument. It may be possible to obtain data that closely approximate the normal distribution through the use of specific instructions on how to distribute ratings. In this study the results of such an effort are evaluated.

Effects of non-normality on statistics and power to detect group differences.

Several studies have concluded that restrictions in score range, as a result of a common response set, result in lower standard deviations and errors of prediction simply because the data are proximally and tightly arranged. Therefore, responses appear more reliable than they may actually be (e.g. Bardo & Yeager, 1982). Consequently, the likelihood of Type I error is increased over that of a normal response distribution. Alternatively, Hui and Triandis (1985) point out that the effect of a wider-than-normal response range on hypothesis test values is that results are more likely to be non-significant than estimates from normally distributed samples. The result is an increase in Type II errors.

A review of the literature of the impact of non-normality on Type I error and power for difference tests in repeated measures ANOVA, as well as on correlations, can be found in recent articles by Keselman and Lix (1995) and Lix, et al. (1996).

Effects of instructions of distributions of responses.

In general, people acquiesce to the demands of the tasks presented in a survey. In numerous studies it has been shown that participants are able to differentiate among and respond differently to different types of instructions. For example, Douglas, McDaniel, and Snell (as cited in Griffith, 1997) found that instructions to fake good and answered honestly on a personality survey presented as part of a selection battery impacted the internal consistency, construct and criterion validities of their measure. Zickar, Rosse, and Levin (as cited in Griffith, 1997) reported changes in rank ordering of applicants when different percentages of "fakers" were introduced into Monte Carlo studies.

Differences in response styles under differing sets of instructions have been demonstrated in survey data from many sources. A confounding factor in this type of inquiry is that survey completion is often conducted in the absence of an administrator. Consequently, researchers must rely on the written instructions to guide respondents through the process. With uniform instructions we hope to reduce the ambiguity of the situation for participants and standardize the reporters' context. In turn, researchers expect to get consistent, reliable, and comparable responses.

Cronbach (1946, 1950) suggested that restriction of response range may be due to learning effects that reduce the variability in responses, lower validities, and heighten estimated reliabilities of the tests. Indeed, more recent work has shown that persons with more experience with a job, coaching, or greater knowledge of a psychological construct are able to manipulate their responses to reflect a greater degree of association with the variable than those less familiar with the concepts (Cronbach, 1946, 1950; Griffith, 1997; Paulhaus, 1991). If people perceive that one form of answer over another is to their advantage they tend to choose the more attractive response. In addition, Cronbach (1946) proposed that scores on ambiguous measures, such as personality or attitude surveys, may be more susceptible to response sets than are more structured situations.

Hypotheses.

The purpose of the present investigation was to examine the effect of completion instructions on univariate and multivariate distributional characteristics and relationships among variables. Instructions a) allowed free-choice allotment of ratings (non-forced instructions) and b) requested the subject to assign a certain number of ratings to either the highest or lowest rating-scale categories (forced instructions).

It was hypothesized that:

H_1: There are significant effects of instructions on the central tendency and dispersion of the ratings.

H_2: There are significant effects of instructions on the shapes of the distributions of the ratings.

H_3: There are significant effects of instructions on the internal consistency of the scale.

Method

Subjects.

A total of 126 participants using the non-forced distributions (NFDs) were obtained from a self-report survey completed in 1996 in a large southern utility company. The comparison sample of 346 forced distributions (FDs) of ratings was collected from 5 companies between 1996 and 1998. Cases were deleted for recording errors and missing responses. Samples of equivalent size (n=115), that contained complete sets of data on the 6-item Group Involvement (GI) scale, were selected from the two instructional-set groups for purposes of comparison.

Procedure

Copies of the two forms of the personality survey had been distributed and collected in previous investigations in companies across the United States. The 1996a version (NFD) had been administered to employees of a large utility company in 1996. The 1996b version (FD) had been used to survey employees in 5 large public and private corporations between 1996 and 1999.

In order to test the hypothesis that distributions of ratings could be affected by instructions for completion, two sets of instructions for a self-report personality survey were presented. One set was a variation on the method of forced distribution (FD) of ratings and the other set asked participants to use the entire width of the scale (NFD). Data from the surveys were analyzed and compared across instruction groups and with characteristics of the normal distribution.

Univariate frequency histograms, skewness and kurtosis coefficients, the Kolmogorov-Smirnov one-sample test, and plots of normal probability were used to examine the data. Marginal distributions and bivariate scatterplots of each pair of items were examined for obvious signs of multivariate non-normality. To assess the impact of instruction set on a relevant statistic, internal consistency (Cronbach alpha), was also estimated on the scale data from each instructional-set group and compared.

Instrument.

The instrument used to collect these data was the Employee Profile© (Somerville and Company, Inc. 1996a, 1996b), a 144-item survey of personality characteristics relevant to work-related contexts. The 1996a version (NFD) contained directions to the respondents to rate themselves on a scale of 1 to 6 points according to how well the statements described their modes of behavior in work-related situations. In addition, the instructions requested that subjects distribute their ratings across the entire 6-point scale. See Appendix A for a sample item and its response format.

The 1996b version (FD) asked participants to rate themselves on a 10-point scale (1 to 10). As well, they were to put at least 20 of the ratings into either of the two lowest categories (1 and 2) and to not put more than 20 of the ratings into either of the two highest categories (9 and 10). A sample item with this format is included in Appendix B.

Raw data were used in the initial analyses to compare the groups. However the differences in the response formats (6-point versus 10-point) that is unrelated to the instructions introduced a confounding variable into the study. In an attempt to control for differences in the number of scale points the two response formats were collapsed into 5-point scales and compared again. The 3 and 4 categories were combined in the 6-point scale and the 10-point scale reduced by pairing adjacent categories (i.e., 1 and 2, 3 and 4, etc.).

The Group Involvement scale consists of 6 items that describe work-related behaviors. Group involvement describes the propensity toward involving one's self in team efforts and to publicly recognize and promote members' contributions. See Figure 1 for the items.

Figure 1.
Items on the Group Involvement scale.

ITEM	STATEMENT
1	Shows a sincere interest in suggestions from members of the work group.
2	Seeks out the special talents and abilities of others to contribute to the quality of the team's product.
3	Encourages people to speak up even when their opinions differ from the opinions of the majority.
4	Develops a feeling of unity and sharing among co-workers.
5	Publicly shares credit for success with those who contributed.
6	Actively promotes the involvement of people having a stake in the outcome of the project or task.

Results

Initial analyses revealed evidence of skewness, kurtosis, and other symptoms of non-normality in both of the samples' response distributions. The FD pattern appears to be a lumpy combination of a bimodal distribution at the extremes with negative skewing in the middle of the rating scale (see Appendix C). The NFD instructional set provided more smoothly distributed data that was, however, peaked and negatively skewed. Response frequency counts and histograms and degrees of skewness and kurtosis were closer to the normal distribution for more of the items in the NFD sample (Table 1) than in the FD (Table 2). NFD responses appeared similar to those of the expected normal distribution in probability plots. Plots of a representative item (Item 1, collapsed and raw data) from the NFD and FD sets are shown in Appendix D.

Table 1.
Skewness and kurtosis in raw and collapsed data in the non-forced (NFD) instruction group.

	Skewness		Kurtosis	
	Raw scale	Collapsed Scale	Raw scale	Collapsed Scale
ITEM 1A	-.016	.721	.057	-.436
ITEM 2A	-.205	.495	-.100	.694
ITEM 3A	-.167	.386	-.568	.342
ITEM 4A	-.296	.430	.178	.632
ITEM 5A	-.369	.267	.678	-.272
ITEM 6A	-.280	.512	.949	1.368

Table 2.
Skewness and kurtosis in raw and collapsed data in the forced (FD) instruction group.

	Skewness		Kurtosis	
	Raw scale	Collapsed Scale	Raw scale	Collapsed Scale
ITEM 1B	-.943	-.962	.007	.055
ITEM 2B	-.557	-.635	-.757	-.641
ITEM 3B	-.324	-.331	-1.381	-1.428
ITEM 4B	-.573	-.663	-.908	-.754
ITEM 5B	-1.380	-1.328	1.475	1.606
ITEM 6B	-.577	-.640	-.894	-.777

Presented below in Table 3 are the comparisons of the data in the collapsed, 5-point scale format. Descriptive statistics for the raw data appear in Appendix E. The rank order of the item means was roughly the same for the two sets and the pairs of item means did not differ significantly from each other at the $p < .05$ level. Standard deviations and standard errors of the means were consistently higher for the FD than for the NFD instructional set data. Standard deviations ranged from 1.052 to 1.377 in the FD group and from .770 to .841 in the NFD set. Standard errors of estimating the means were more consistent across items in the NFD data (.072 to .078) than were the FD responses (.098 to .128). The differences observed here in the variability of items are directly related to the differences in the estimates of scale reliability (internal consistency) reported later in this paper.

Table 3.
Means and standard deviations of the collapsed scales by instruction group.

	Non-Forced Distribution Collapsed Scale			Forced Distribution Collapsed Scale		
	Mean	Standard Deviation	Standard Error of the Mean	Mean	Standard Deviation	Standard Error of the Mean
Item 1	3.504	.799	.074	3.470	1.157	.108
Item 2	3.217	.781	.073	3.261	1.271	.119

Item 3	3.130	.800	.075	2.878	1.377	.128
Item 4	3.235	.841	.078	3.130	1.196	.111
Item 5	3.617	.833	.078	3.774	1.052	.098
Item 6	3.287	.770	.072	3.139	1.235	.115

Bivariate scatterplots of the NFD data followed the expected, elliptical patterning of multivariate normal distributions to a greater degree than did the FD data. In contrast, the FD scatterplots appeared to reflect the skewed and bimodal nature of the univariate distributions exhibited in the item's categorical frequency histograms.

Visual inspection of the normal probability plots revealed a distinct trend for greater deviation from normality in the FD distributions across all item comparisons for both the raw and collapsed response data. (See the example in Appendix D.) The NFD data appeared to closely approximate the linear function and the FD data did not.

The Kolmorogov-Smirnov one-sample test for normality (KS) is used for ordinal or higher scale data and the d-statistic is calculated based on the difference in the observed cumulative distribution and the hypothesized, in this case normal, distribution. For the purposes of this study, the region of rejection was determined a priori at the $p < .05$ level. When the KS d-value is statistically significant, we reject the hypothesis that the observed data follow the hypothesized, normal distribution. All of the items' rating distributions were significantly different from the normal distribution (Table 4).

Table 4.
Kolmorogov-Smirnov test for approximation of the normal distribution.

	Raw Data		Collapsed Data	
	Non-Forced Distribution	Forced Distribution	Non-Forced Distribution	Forced Distribution
	d =	d =	d =	d =
Item 1	.275	.230	.353	.329
Item 2	.244	.172	.340	.250
Item 3	.232	.196	.304	.262
Item 4	.265	.205	.340	.279
Item 5	.232	.244	.301	.342
Item 6	.266	.203	.376	.270

Note: Distributions were all significantly ($p < .05$) different from normal distribution.

The scale statistics demonstrated that there were substantial differences between the responses to the instructional sets. In the raw data Cronbach alpha for the GI scale in the NFD condition was much higher than that of the FD group in the raw ($\alpha =$ .767 versus .453) and collapsed ($\alpha =$ .770 versus .397) data sets. Average inter-item correlations were higher in the NFD (raw $r = .357$, collapsed $r = .360$) than in the FD group (raw $r = .125$, collapsed $r = .101$). Skewness of the total scale was greater in the NFD data and kurtosis was greater in the FD condition (see Table 5). Results in the collapsed data mirrored those of the raw data. In the case of the NFD sample the degree of skewness and kurtosis increased as a result of condensing the scale and the distribution switched from being platykurtic to leptokurtotic. Skewness decreased in the FD data when the categories were collapsed and kurtosis remained around zero.

Table 5.
Scale statistics.

	Raw Data				Collapsed Data			
	Cronbach's alpha	Average inter-item correlation	Skewness	Kurtosis	Cronbach's alpha	Average inter-item correlation	Skewness	Kurtosis
NON-FORCED DISTRIBUTION	.767	.357	.130	-.108	.770	.360	.607	.241
FORCED DISTRIBUTION	.453	.125	-.228	-.030	.397	.101	-.130	.030

Discussion

It was hypothesized that there would be effects of instructions on the central tendency and dispersion of the ratings. Although the collapsed means did not differ across instructional sets the variability in the ratings was higher with the FD directions. Standard deviations and standard errors of the mean were higher for every item answered under the FD instructions.

Support for the second hypothesis, that there would be effects of instructions on the shapes of the distributions of the ratings, was found in this study. The shapes of the distributions were quite different across sets and, apparently, neither set of respondent instructions elicited normally distributed data. However, the instructional set that allowed free-choice distributions of personality survey item ratings seemed to provide raw score data that more closely approximated the normal distribution than did the FD set. In the frequency histograms the FD distribution appeared much lumpier than did the NFD data. In tests of skewness, kurtosis, and normality as well as in normal probability plots the NFD data were usually closer to the normal than were the FD responses.

The same phenomenon was observed in the 5-point, collapsed scale distributions. FD ratings conformed to distributions that differed from the normal to a greater degree than ratings obtained with the NFD instructional set. Skewness and kurtosis were higher and collapsed format NFD data were less normally distributed, according to the KS test statistics, than were their raw counterparts. However, in the FD data, collapsing the scale resulted in some scales becoming more normal and others becoming less normal in their distribution characteristics.

Hypothesis 3 stated that there would be an effect of instructions on the internal consistency of the scale. This hypothesis was also supported by the results of the study. It was found that the reliability of the Group Involvement scale was underestimated by the FD data when compared with the NFD. The forced-ratings distribution method appeared to impact the skewness and kurtosis of the scale and the item intercorrelations, as well.

Researchers (e.g., Bardo & Yeager, 1982; Townsend & Ashby, 1984) have found rating scales more susceptible to response bias than are unlabeled or distally labeled numerical scales. The 6-point scale was completely labeled and 10-point had only distal anchors to describe it. As mentioned by Bardo and Yeager (1982), researchers have found that the less ambiguous the rating scale labels, the less variability due to error of interpretation is contained in responses. The differences in the number of scale points, rating scale descriptors, and formats of the response categories are confounding influences on the study results.

Caution using collapsed scales is warranted by these results. In most, but not all cases, skewness and kurtosis in the combined rating categories were different from the raw data. The attempt to transform the data, as Townsend and Ashby (1984) and others have warned, resulted in messier distributions that were even less amenable to the use of inferential statistics than the original data.

The contribution of the present investigation to knowledge in the field of measurement and survey design is twofold. First, two methods of distributing self-report ratings, one of which appears to have remained unreported in the literature, were analyzed and compared. The hypothesis that the instructions would impact the normality of distributions in self-report surveys was supported. It would appear that asking participants to use the range of their response options is a more effective way of eliciting normally distributed data than is asking them to assign a proportion of their responses to the extreme categories.

The results also serve as a cautionary note to survey designers and consumers. When we use self-report measures, as most survey data are, the distributions of data on our variables may not be normal. In this case, forcing categorical decisions onto a rating scale along with decisions that were to be made on a continuum created rather messy, non-normal distributions. And the FD instructions appeared to have an impact on statistics derived from the sample under these conditions. Certainty in conclusions based on these data was compromised by the violations of assumptions underlying tests of differences and associations. It may be less than prudent to assume that true error rates are equal to the selected alpha level for testing hypotheses on sample data obtained under these conditions.

Another interesting result of the study was finding evidence that even procedures as simple as combining categories on rating scales can affect the results and interpretations of survey data. Researchers would be well advised to consider alternative ways to compare data gathered using different response formats. Then comparative analyses of the distributional characteristics of the transformed with the raw data sets should be made and the best fitting transform selected before results produced from further analyses are deemed equivalent.

As the opening quote by Townsend and Ashby (1984) suggests, we need to make sure that our conclusions will be the same regardless of the rulers we use to measure our ideas. Survey researchers face tremendous odds against finding high-precision, equivalent, meaningful methods of measurement because of factors such as the impact of instructional sets or response formats in surveys of characteristics, attitudes, and behaviors. Reports of investigations into alternative methods of measurement, such as item response theory, help to direct the search for more accurate ways to compare and evaluate people.

References

Bardo, J. W., & Yeager, S. J. (1982). Consistency of response style across types of response formats. Perceptual and Motor Skills, 55, 307-310.

Barrick, M., & Mount, M. ((1996). Effects of impression management and self-deception on the predictive validity of personality constructs. Journal of Applied Psychology, 81, 261-272.

Cronbach, L. J. (1950). Further evidence on response sets and test design. Educational and Psychological Measurement, 10, 3-31.

Edwards, A. L. (1953). The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. Journal of Applied Psychology, 37, 90-93.

Furnham, A. (1986). Response bias, social desirability and dissimulation. Personality and Individual Differences, 7, 385-400.

Griffith, R. (1997). Faking of noncognitive selection devices: Red herring is hard to swallow. Unpublished doctoral dissertation, University of Akron, Ohio.

Hancock, G., & Klockars, A. (1996). The quest for alpha: Developments in multiple comparison procedures in the quarter century since Games (1971). Review of Educational Research, 66, 269-306.

Hough, L., Eaton, N., Dunnette, M., Kamp, J., & McCloy, R. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. Journal of Applied Psychology, 75, 581-595.

Hui, C. H., & Triandis, H. C. (1985). The instability of response sets. Public Opinion Quarterly, 49, 253-260.

Judd, C., McClelland, G., & Culhane, S. (1995). Data analysis: Continuing issues in the everyday analysis of psychological data. Annual Review of Psychology, 46, 433-465.

Keselman, H. & Lix, L. (1995). Improved repeated measures stepwise multiple comparison procedures. Journal of Educational and Behavioral Statistics, 20, 83-99.

Lix, L., Keselman, J., & Keselman, H. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test.
Review of Educational Research, 66, 579-619.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. Psychological Bulletin, 105, 156-166.

Ones, D., & Viswesvaran, C. (1998). The effects of social desirability and faking on personality and integrity assessment for personnel selection. Human Performance, 11, 245-269.

Ones, D., Viswesvaran, C., & Reiss, A. (1996). Role of social desirability in personality testing for personnel selection: The red herring. Journal of Applied Psychology, 81, 660-679.

Paulhaus, D. (1984). Two-component models of socially desirable responding. Journal of Personality and Social Psychology, 46, 598-609.

Paulhaus, D. (1986). Self-Deception and impression management in test responses. In A. Angleitner & J. Wiggins (Eds.), Personality Assessment via Questionnaires (pp. 143-165). Berlin: Springer-Verlag.

Townsend, J., & Ashby, F. (1984). Measurement scales and statistics: The misconception misconceived. Psychological Bulletin, 96, 394-401.

APPENDIX A

Non-forced distribution instructions to respondents on the Employee Profile.

The Employee Profile Survey was designed to assist in identifying an individual's most and least prominent work behaviors and characteristics. Your ratings should be based on how frequently these behaviors occur or how characteristic that behavior is of you.

Each person should exhibit some traits that are obvious to those around them. The behaviors demonstrated most frequently should be rated 5 or 6. This will identify your most prominent behavioral features.

It would be likely that a person would not demonstrate some of these behaviors very often. Rate some of the items 1 or 2 to identify your least prominent behavioral characteristics.

Respond to the remaining items with 3 or 4 ratings depending upon how often or how characteristic you think these behaviors are of you.

Almost Never ① Not at All Characteristic	Seldom or Once in a while ② Slightly	Regularly but Not often. ③ Moderately Characteristic	Fairly Often ④ Characteristic	Very Frequently ⑤ Very Characteristic	Almost Always ⑥ Extremely

1A. Shows a sincere interest in suggestions from members of the work group. ①②③④⑤⑥

APPENDIX B

Forced distribution instructions to respondents on the Employee Profile.

Read each item and determine how characteristic of you the behavior associated with the item is using the "1" to "10" scale. In order to ensure that your most and least characteristic traits are identified, rate at least twenty (20) items "1" or "2", and no more than twenty (20) items at "9 or 10." You may find that putting 5 items in the "1" or "2" and 5 in the "9" or "10" ranges on EACH page is easiest way to accomplish this. Rate the remaining items within the range "3" and "8" depending upon your assessment of how characteristic of you each item is.

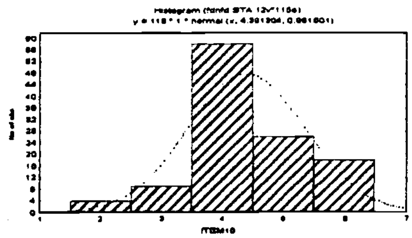
Less Characteristic	①②	③④⑤⑥⑦⑧	⑨⑩	**More Characteristic**

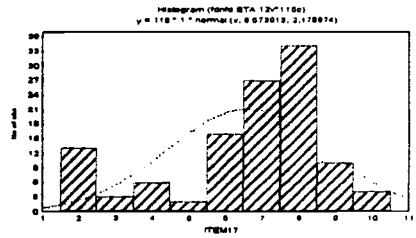
1B Shows a sincere interest in suggestions from members of the work group ①② ③④⑤⑥⑦⑧ ⑨⑩

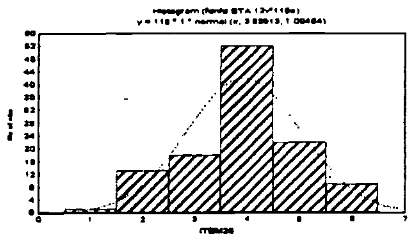
APPENDIX C

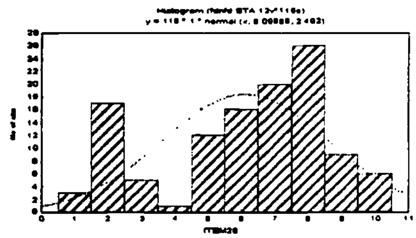
Frequency histograms for items 1-6 (raw data)

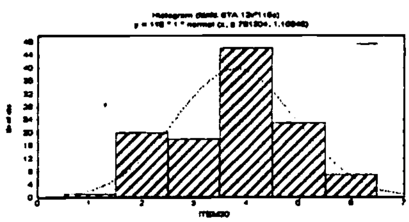
6-point scale	10-point scale

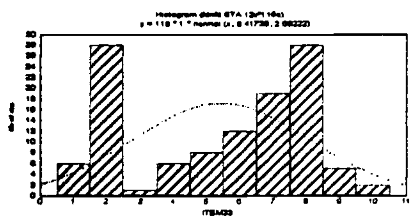


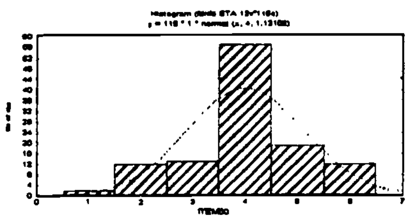


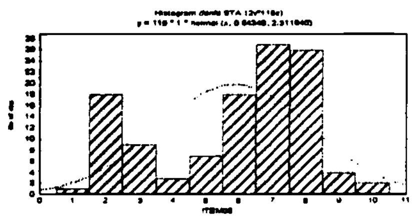


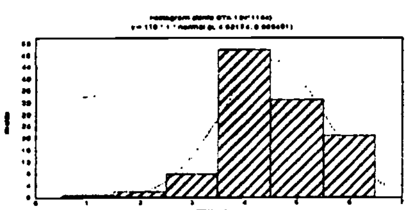


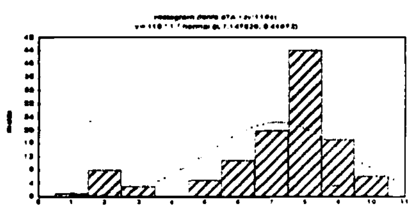


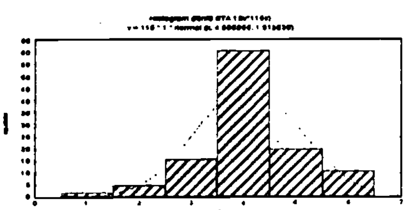


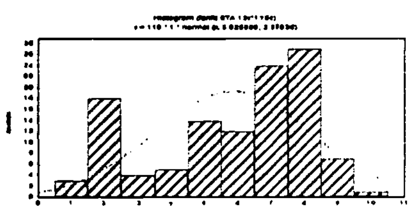






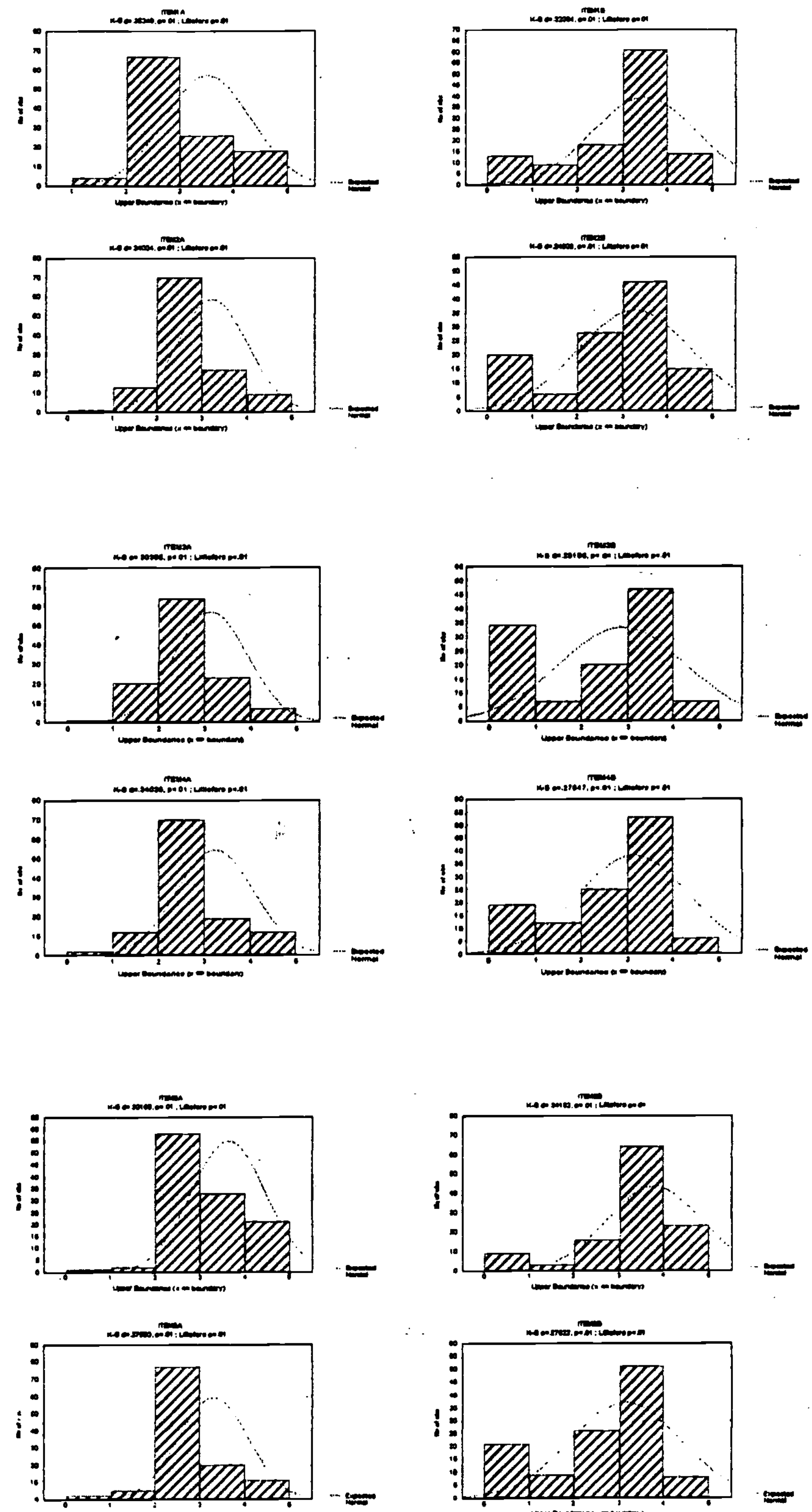






Frequency histograms for items 1-6 (collapsed data)

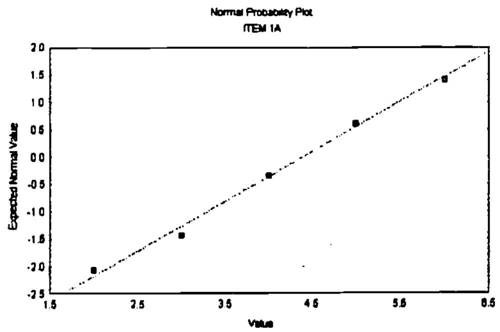
collapsed 6-point scale | collapsed 10-point scale



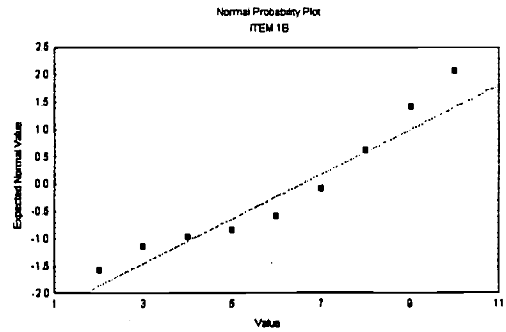
APPENDIX D

Normal probability plots for Item 1

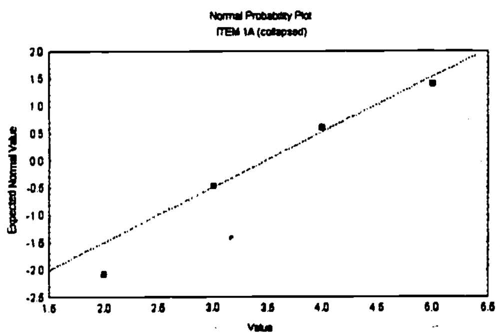
a. Raw data, NFD set.



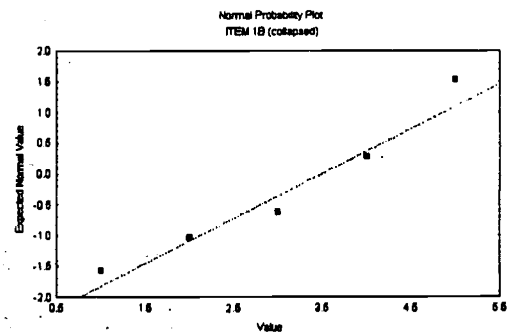
b. Raw data, FD set.



c. Collapsed data, NFD set.



d. Collapsed data, FD set.



BEST COPY AVAILABLE

APPENDIX E

Raw scale means and dispersion.

Non-forced distribution; 6-point scale (1-6)

	Mean	Standard Deviation	Standard Error
ITEM 1A	4.391	0.961	.090
ITEM 2A	3.939	1.095	.102
ITEM 3A	3.791	1.159	.108
ITEM 4A	4.000	1.132	.105
ITEM 5A	4.522	0.985	.092
ITEM 6A	4.087	1.014	.095

Forced distribution; 10-point scales (1-10)

	Mean	Standard Deviation	Standard Error
ITEM 1B	6.574	2.177	.203
ITEM 2B	6.096	2.492	.232
ITEM 3B	5.417	2.662	.248
ITEM 4B	5.843	2.312	.216
ITEM 5B	7.148	2.049	.191
ITEM 6B	5.826	2.370	.221

The Impact of Instructional Set on Distributions of Self-Report Ratings on a Survey of Personality Characteristics.

Paper presented at AERA
New Orleans, 2000
Jo-Anne McDonald
and Lisa Hall
University of Denver
Denver, Colorado

The impact of instructional set on distributions of self-report ratings on a survey of personality characteristics.

Introduction

"A statement about an empirical system is meaningful only when it is scale independent, that is, only when it is true on all of the permissible numeric scales".

-Townsend and Ashby (1984, pp. 399)

Researchers seeking to evaluate the impact of program variables on performance criteria often use survey and self-report data to do so. But often neglected is an examination of evidence supporting or conflicting with the assumptions underlying our theories and their measurement.

Statistical tests frequently assume (as do statisticians) that the level of a psychological variable is normally distributed in the populations from which samples are drawn. While most of the limited evidence points to the relatively slight impact of the violation of the normality assumption on power and significance tests (Lix, Keselman, & Keselman, 1996) as Townsend and Ashby (1984) point out, further studies are needed before we can assume these findings generalize to all applications. Indeed, researchers have found differences in results of a variety of inferential statistics in studies involving manipulations of distribution parameters, response formats, and instruction sets (e.g., Hancock & Klockars, 1996; Hough, Eaton, Dunnette, Kamp, & McCloy, 1990; Keselman & Lix, 1995). However, many of these studies used simulated data sets that conformed to a regular, known distribution.

Micceri (1989) found that, in a meta-analysis of distributional qualities in behavioral science data sets, none of over 400 data sets that he assessed passed his criteria for normality. The distribution of scores on psychometric measures, such as ability or attitude scales and achievement measures, were universally asymmetrical and "lumpy". In addition, Micceri suggested that for most data sets, non-parametric analysis might provide more accurate information and more powerful tests in social research applications.

The lack of normality in general response patterns may be compounded by social desirability and faking response styles on self-report inventories of personality or surveys of attitudes. One component of social desirability has been termed "impression management". Paulhaus (1984, 1986) believed self-report to be one way that people use to control what others think of them. They estimate situational demands and modify their behavior to enhance the probability of a preferred outcome. This parallels earlier work in social desirability by Edwards (1953). He found a high correlation (r = .87) between increase in an item's social desirability and the tendency for people to rate themselves higher on the item.

In a review of the social desirability, response bias, and faking literature, Furnham (1986) concluded that personality and adjustment measures are affected to a greater extent than other types of questionnaires. More recent reports indicate that people do fake responses and that self-report measures are particularly susceptible to faking "good" as well as faking "bad" response styles or the tendency to present one's self in a beneficial light (e.g., Barrick & Mount, 1996; Griffith, 1997; Hough, et al., 1990). However, other researchers have found that social desirability did not decrease the criterion-related validity of personality predictors of job performance (e.g., Hough, et al., 1990; Ones, Viswesvaran, & Reiss, 1996).

Another challenge for survey researchers is to assess the impact of response formats on results. Rating scales appear to be particularly susceptible to response sets more so than distally anchored or purely numeric response scales (e.g., Bardo & Yeager, 1982; Cronbach, 1950). This condition appears to hold across scale formats and number of rating categories.

Distributions of variables of interest to social scientists are often restricted, skewed, and leptokurtotic, in other words, not normally distributed. What can be done about it and how do we avoid the problem?

One solution researchers have devised is to mathematically transform the data. However, the "best" solution to the problem may be elusive. In addition, interpretations of the results are complicated by the altered state of the variables (e.g., Judd, et al., 1995; Townsend & Ashby, 1984). A second solution is to provide the opportunity for respondents to select between two statements of equal desirability. A third solution is to make the intention of the item more obscure. obscurity makes it harder for people to intentionally distort their responses. A fourth solution is to warn the respondents that their answers will be evaluated for accurateness using outside criteria. A fifth method is to include scales that detect response patterns associated with dissimulation.

Yet another approach is to circumvent the problem by manipulating the instructions for completion of an instrument. It may be possible to obtain data that closely approximate the normal distribution through the use of specific instructions on how to distribute ratings. In this study the results of such an effort are evaluated.

Effects of non-normality on statistics and power to detect group differences.

Several studies have concluded that restrictions in score range, as a result of a common response set, result in lower standard deviations and errors of prediction simply because the data are proximally and tightly arranged. Therefore, responses appear more reliable than they may actually be (e.g. Bardo & Yeager, 1982). Consequently, the likelihood of Type I error is increased over that of a normal response distribution. Alternatively, Hui and Triandis (1985) point out that the effect of a wider-than-normal response range on hypothesis test values is that results are more likely to be non-significant than estimates from normally distributed samples. The result is an increase in Type II errors.

A review of the literature of the impact of non-normality on Type I error and power for difference tests in repeated measures ANOVA, as well as on correlations, can be found in recent articles by Keselman and Lix (1995) and Lix, et al. (1996).

Effects of instructions of distributions of responses.

In general, people acquiesce to the demands of the tasks presented in a survey. In numerous studies it has been shown that participants are able to differentiate among and respond differently to different types of instructions. For example, Douglas, McDaniel, and Snell (as cited in Griffith, 1997) found that instructions to fake good and answered honestly on a personality survey presented as part of a selection battery impacted the internal consistency, construct and criterion validities of their measure. Zickar, Rosse, and Levin (as cited in Griffith, 1997) reported changes in rank ordering of applicants when different percentages of "fakers" were introduced into Monte Carlo studies.

Differences in response styles under differing sets of instructions have been demonstrated in survey data from many sources. A confounding factor in this type of inquiry is that survey completion is often conducted in the absence of an administrator. Consequently, researchers must rely on the written instructions to guide respondents through the process. With uniform instructions we hope to reduce the ambiguity of the situation for participants and standardize the reporters' context. In turn, researchers expect to get consistent, reliable, and comparable responses.

Cronbach (1946, 1950) suggested that restriction of response range may be due to learning effects that reduce the variability in responses, lower validities, and heighten estimated reliabilities of the tests. Indeed, more recent work has shown that persons with more experience with a job, coaching, or greater knowledge of a psychological construct are able to manipulate their responses to reflect a greater degree of association with the variable than those less familiar with the concepts (Cronbach, 1946, 1950; Griffith, 1997; Paulhaus, 1991). If people perceive that one form of answer over another is to their advantage they tend to choose the more attractive response. In addition, Cronbach (1946) proposed that scores on ambiguous measures, such as personality or attitude surveys, may be more susceptible to response sets than are more structured situations.

Hypotheses.

The purpose of the present investigation was to examine the effect of completion instructions on univariate and multivariate distributional characteristics and relationships among variables. Instructions a) allowed free-choice allotment of ratings (non-forced instructions) and b) requested the subject to assign a certain number of ratings to either the highest or lowest rating-scale categories (forced instructions).

It was hypothesized that:

H_1: There are significant effects of instructions on the central tendency and dispersion of the ratings.

H_2: There are significant effects of instructions on the shapes of the distributions of the ratings.

H_3: There are significant effects of instructions on the internal consistency of the scale.

Method

Subjects.

A total of 126 participants using the non-forced distributions (NFDs) were obtained from a self-report survey completed in 1996 in a large southern utility company. The comparison sample of 346 forced distributions (FDs) of ratings was collected from 5 companies between 1996 and 1998. Cases were deleted for recording errors and missing responses. Samples of equivalent size (n=115), that contained complete sets of data on the 6-item Group Involvement (GI) scale, were selected from the two instructional-set groups for purposes of comparison.

Procedure

Copies of the two forms of the personality survey had been distributed and collected in previous investigations in companies across the United States. The 1996a version (NFD) had been administered to employees of a large utility company in 1996. The 1996b version (FD) had been used to survey employees in 5 large public and private corporations between 1996 and 1999.

In order to test the hypothesis that distributions of ratings could be affected by instructions for completion, two sets of instructions for a self-report personality survey were presented. One set was a variation on the method of forced distribution (FD) of ratings and the other set asked participants to use the entire width of the scale (NFD). Data from the surveys were analyzed and compared across instruction groups and with characteristics of the normal distribution.

Univariate frequency histograms, skewness and kurtosis coefficients, the Kolmogorov-Smirnov one-sample test, and plots of normal probability were used to examine the data. Marginal distributions and bivariate scatterplots of each pair of items were examined for obvious signs of multivariate non-normality. To assess the impact of instruction set on a relevant statistic, internal consistency (Cronbach alpha), was also estimated on the scale data from each instructional-set group and compared.

Instrument.

The instrument used to collect these data was the Employee Profile© (Somerville and Company, Inc. 1996a, 1996b), a 144-item survey of personality characteristics relevant to work-related contexts. The 1996a version (NFD) contained directions to the respondents to rate themselves on a scale of 1 to 6 points according to how well the statements described their modes of behavior in work-related situations. In addition, the instructions requested that subjects distribute their ratings across the entire 6-point scale. See Appendix A for a sample item and its response format.

The 1996b version (FD) asked participants to rate themselves on a 10-point scale (1 to 10). As well, they were to put at least 20 of the ratings into either of the two lowest categories (1 and 2) and to not put more than 20 of the ratings into either of the two highest categories (9 and 10). A sample item with this format is included in Appendix B.

Raw data were used in the initial analyses to compare the groups. However the differences in the response formats (6-point versus 10-point) that is unrelated to the instructions introduced a confounding variable into the study. In an attempt to control for differences in the number of scale points the two response formats were collapsed into 5-point scales and compared again. The 3 and 4 categories were combined in the 6-point scale and the 10-point scale reduced by pairing adjacent categories (i.e., 1 and 2, 3 and 4, etc.).

The Group Involvement scale consists of 6 items that describe work-related behaviors. Group involvement describes the propensity toward involving one's self in team efforts and to publicly recognize and promote members' contributions. See Figure 1 for the items.

Figure 1.
Items on the Group Involvement scale.

ITEM	STATEMENT
1	Shows a sincere interest in suggestions from members of the work group.
2	Seeks out the special talents and abilities of others to contribute to the quality of the team's product.
3	Encourages people to speak up even when their opinions differ from the opinions of the majority.
4	Develops a feeling of unity and sharing among co-workers.
5	Publicly shares credit for success with those who contributed.
6	Actively promotes the involvement of people having a stake in the outcome of the project or task.

Results

Initial analyses revealed evidence of skewness, kurtosis, and other symptoms of non-normality in both of the samples' response distributions. The FD pattern appears to be a lumpy combination of a bimodal distribution at the extremes with negative skewing in the middle of the rating scale (see Appendix C). The NFD instructional set provided more smoothly distributed data that was, however, peaked and negatively skewed. Response frequency counts and histograms and degrees of skewness and kurtosis were closer to the normal distribution for more of the items in the NFD sample (Table 1) than in the FD (Table 2). NFD responses appeared similar to those of the expected normal distribution in probability plots. Plots of a representative item (Item 1, collapsed and raw data) from the NFD and FD sets are shown in Appendix D.

Table 1.
Skewness and kurtosis in raw and collapsed data in the non-forced (NFD) instruction group.

	Skewness		Kurtosis	
	Raw scale	Collapsed Scale	Raw scale	Collapsed Scale
ITEM 1A	-.016	.721	.057	-.436
ITEM 2A	-.205	.495	-.100	.694
ITEM 3A	-.167	.386	-.568	.342
ITEM 4A	-.296	.430	.178	.632
ITEM 5A	-.369	.267	.678	-.272
ITEM 6A	-.280	.512	.949	1.368

Table 2.
Skewness and kurtosis in raw and collapsed data in the forced (FD) instruction group.

	Skewness		Kurtosis	
	Raw scale	Collapsed Scale	Raw scale	Collapsed Scale
ITEM 1B	-.943	-.962	.007	.055
ITEM 2B	-.557	-.635	-.757	-.641
ITEM 3B	-.324	-.331	-1.381	-1.428
ITEM 4B	-.573	-.663	-.908	-.754
ITEM 5B	-1.380	-1.328	1.475	1.606
ITEM 6B	-.577	-.640	-.894	-.777

Presented below in Table 3 are the comparisons of the data in the collapsed, 5-point scale format. Descriptive statistics for the raw data appear in Appendix E. The rank order of the item means was roughly the same for the two sets and the pairs of item means did not differ significantly from each other at the $p < .05$ level. Standard deviations and standard errors of the means were consistently higher for the FD than for the NFD instructional set data. Standard deviations ranged from 1.052 to 1.377 in the FD group and from .770 to .841 in the NFD set. Standard errors of estimating the means were more consistent across items in the NFD data (.072 to .078) than were the FD responses (.098 to .128). The differences observed here in the variability of items are directly related to the differences in the estimates of scale reliability (internal consistency) reported later in this paper.

Table 3.
Means and standard deviations of the collapsed scales by instruction group.

	Non-Forced Distribution Collapsed Scale			Forced Distribution Collapsed Scale		
	Mean	Standard Deviation	Standard Error of the Mean	Mean	Standard Deviation	Standard Error of the Mean
Item 1	3.504	.799	.074	3.470	1.157	.108
Item 2	3.217	.781	.073	3.261	1.271	.119

Item 3	3.130	.800	.075	2.878	1.377	.128
Item 4	3.235	.841	.078	3.130	1.196	.111
Item 5	3.617	.833	.078	3.774	1.052	.098
Item 6	3.287	.770	.072	3.139	1.235	.115

Bivariate scatterplots of the NFD data followed the expected, elliptical patterning of multivariate normal distributions to a greater degree than did the FD data. In contrast, the FD scatterplots appeared to reflect the skewed and bimodal nature of the univariate distributions exhibited in the item's categorical frequency histograms.

Visual inspection of the normal probability plots revealed a distinct trend for greater deviation from normality in the FD distributions across all item comparisons for both the raw and collapsed response data. (See the example in Appendix D.) The NFD data appeared to closely approximate the linear function and the FD data did not.

The Kolmorogov-Smirnov one-sample test for normality (KS) is used for ordinal or higher scale data and the d-statistic is calculated based on the difference in the observed cumulative distribution and the hypothesized, in this case normal, distribution. For the purposes of this study, the region of rejection was determined a priori at the $p < .05$ level. When the KS d-value is statistically significant, we reject the hypothesis that the observed data follow the hypothesized, normal distribution. All of the items' rating distributions were significantly different from the normal distribution (Table 4).

Table 4.
Kolmorogov-Smirnov test for approximation of the normal distribution.

	Raw Data		Collapsed Data	
	Non-Forced Distribution	Forced Distribution	Non-Forced Distribution	Forced Distribution
	d =	d =	d =	d =
Item 1	.275	.230	.353	.329
Item 2	.244	.172	.340	.250
Item 3	.232	.196	.304	.262
Item 4	.265	.205	.340	.279
Item 5	.232	.244	.301	.342
Item 6	.266	.203	.376	.270

Note: Distributions were all significantly ($p < .05$) different from normal distribution.

The scale statistics demonstrated that there were substantial differences between the responses to the instructional sets. In the raw data Cronbach alpha for the GI scale in the NFD condition was much higher than that of the FD group in the raw (α = .767 versus .453) and collapsed (α = .770 versus .397) data sets. Average inter-item correlations were higher in the NFD (raw r = .357, collapsed r = .360) than in the FD group (raw r = .125, collapsed r = .101). Skewness of the total scale was greater in the NFD data and kurtosis was greater in the FD condition (see Table 5). Results in the collapsed data mirrored those of the raw data. In the case of the NFD sample the degree of skewness and kurtosis increased as a result of condensing the scale and the distribution switched from being platykurtic to leptokurtotic. Skewness decreased in the FD data when the categories were collapsed and kurtosis remained around zero.

Table 5.
Scale statistics.

	Raw Data				Collapsed Data			
	Cronbach's alpha	Average inter-item correlation	Skewness	Kurtosis	Cronbach's alpha	Average inter-item correlation	Skewness	Kurtosis
NON-FORCED DISTRIBUTION	.767	.357	.130	-.108	.770	.360	.607	.241
FORCED DISTRIBUTION	.453	.125	-.228	-.030	.397	.101	-.130	.030

Discussion

It was hypothesized that there would be effects of instructions on the central tendency and dispersion of the ratings. Although the collapsed means did not differ across instructional sets the variability in the ratings was higher with the FD directions. Standard deviations and standard errors of the mean were higher for every item answered under the FD instructions.

Support for the second hypothesis, that there would be effects of instructions on the shapes of the distributions of the ratings, was found in this study. The shapes of the distributions were quite different across sets and, apparently, neither set of respondent instructions elicited normally distributed data. However, the instructional set that allowed free-choice distributions of personality survey item ratings seemed to provide raw score data that more closely approximated the normal distribution than did the FD set. In the frequency histograms the FD distribution appeared much lumpier than did the NFD data. In tests of skewness, kurtosis, and normality as well as in normal probability plots the NFD data were usually closer to the normal than were the FD responses.

The same phenomenon was observed in the 5-point, collapsed scale distributions. FD ratings conformed to distributions that differed from the normal to a greater degree than ratings obtained with the NFD instructional set. Skewness and kurtosis were higher and collapsed format NFD data were less normally distributed, according to the KS test statistics, than were their raw counterparts. However, in the FD data, collapsing the scale resulted in some scales becoming more normal and others becoming less normal in their distribution characteristics.

Hypothesis 3 stated that there would be an effect of instructions on the internal consistency of the scale. This hypothesis was also supported by the results of the study. It was found that the reliability of the Group Involvement scale was underestimated by the FD data when compared with the NFD. The forced-ratings distribution method appeared to impact the skewness and kurtosis of the scale and the item intercorrelations, as well.

Researchers (e.g., Bardo & Yeager, 1982; Townsend & Ashby, 1984) have found rating scales more susceptible to response bias than are unlabeled or distally labeled numerical scales. The 6-point scale was completely labeled and 10-point had only distal anchors to describe it. As mentioned by Bardo and Yeager (1982), researchers have found that the less ambiguous the rating scale labels, the less variability due to error of interpretation is contained in responses. The differences in the number of scale points, rating scale descriptors, and formats of the response categories are confounding influences on the study results.

Caution using collapsed scales is warranted by these results. In most, but not all cases, skewness and kurtosis in the combined rating categories were different from the raw data. The attempt to transform the data, as Townsend and Ashby (1984) and others have warned, resulted in messier distributions that were even less amenable to the use of inferential statistics than the original data.

The contribution of the present investigation to knowledge in the field of measurement and survey design is twofold. First, two methods of distributing self-report ratings, one of which appears to have remained unreported in the literature, were analyzed and compared. The hypothesis that the instructions would impact the normality of distributions in self-report surveys was supported. It would appear that asking participants to use the range of their response options is a more effective way of eliciting normally distributed data than is asking them to assign a proportion of their responses to the extreme categories.

The results also serve as a cautionary note to survey designers and consumers. When we use self-report measures, as most survey data are, the distributions of data on our variables may not be normal. In this case, forcing categorical decisions onto a rating scale along with decisions that were to be made on a continuum created rather messy, non-normal distributions. And the FD instructions appeared to have an impact on statistics derived from the sample under these conditions. Certainty in conclusions based on these data was compromised by the violations of assumptions underlying tests of differences and associations. It may be less than prudent to assume that true error rates are equal to the selected alpha level for testing hypotheses on sample data obtained under these conditions.

Another interesting result of the study was finding evidence that even procedures as simple as combining categories on rating scales can affect the results and interpretations of survey data. Researchers would be well advised to consider alternative ways to compare data gathered using different response formats. Then comparative analyses of the distributional characteristics of the transformed with the raw data sets should be made and the best fitting transform selected before results produced from further analyses are deemed equivalent.

As the opening quote by Townsend and Ashby (1984) suggests, we need to make sure that our conclusions will be the same regardless of the rulers we use to measure our ideas. Survey researchers face tremendous odds against finding high-precision, equivalent, meaningful methods of measurement because of factors such as the impact of instructional sets or response formats in surveys of characteristics, attitudes, and behaviors. Reports of investigations into alternative methods of measurement, such as item response theory, help to direct the search for more accurate ways to compare and evaluate people.

References

Bardo, J. W., & Yeager, S. J. (1982). Consistency of response style across types of response formats. Perceptual and Motor Skills, 55, 307-310.

Barrick, M., & Mount, M. ((1996). Effects of impression management and self-deception on the predictive validity of personality constructs. Journal of Applied Psychology, 81, 261-272.

Cronbach, L. J. (1950). Further evidence on response sets and test design. Educational and Psychological Measurement, 10, 3-31.

Edwards, A. L. (1953). The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. Journal of Applied Psychology, 37, 90-93.

Furnham, A. (1986). Response bias, social desirability and dissimulation. Personality and Individual Differences, 7, 385-400.

Griffith, R. (1997). Faking of noncognitive selection devices: Red herring is hard to swallow. Unpublished doctoral dissertation, University of Akron, Ohio.

Hancock, G., & Klockars, A. (1996). The quest for alpha: Developments in multiple comparison procedures in the quarter century since Games (1971). Review of Educational Research, 66, 269-306.

Hough, L., Eaton, N., Dunnette, M., Kamp, J., & McCloy, R. (1990). Criterion-related validities of personality constructs and the effect of response distortion on those validities. Journal of Applied Psychology, 75, 581-595.

Hui, C. H., & Triandis, H. C. (1985). The instability of response sets. Public Opinion Quarterly, 49, 253-260.

Judd, C., McClelland, G., & Culhane, S. (1995). Data analysis: Continuing issues in the everyday analysis of psychological data. Annual Review of Psychology, 46, 433-465.

Keselman, H. & Lix, L. (1995). Improved repeated measures stepwise multiple comparison procedures. Journal of Educational and Behavioral Statistics, 20, 83-99.

Lix, L., Keselman, J., & Keselman, H. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. Review of Educational Research, 66, 579-619.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. Psychological Bulletin, 105, 156-166.

Ones, D., & Viswesvaran, C. (1998). The effects of social desirability and faking on personality and integrity assessment for personnel selection. Human Performance, 11, 245-269.

Ones, D., Viswesvaran, C., & Reiss, A. (1996). Role of social desirability in personality testing for personnel selection: The red herring. Journal of Applied Psychology, 81, 660-679.

Paulhaus, D. (1984). Two-component models of socially desirable responding. Journal of Personality and Social Psychology, 46, 598-609.

Paulhaus, D. (1986). Self-Deception and impression management in test responses. In A. Angleitner & J. Wiggins (Eds.), Personality Assessment via Questionnaires (pp. 143-165). Berlin: Springer-Verlag.

Townsend, J., & Ashby, F. (1984). Measurement scales and statistics: The misconception misconceived. Psychological Bulletin, 96, 394-401.

APPENDIX A

Non-forced distribution instructions to respondents on the Employee Profile.

The Employee Profile Survey was designed to assist in identifying an individual's most and least prominent work behaviors and characteristics. Your ratings should be based on how frequently these behaviors occur or how characteristic that behavior is of you.

Each person should exhibit some traits that are obvious to those around them. The behaviors demonstrated most frequently should be rated 5 or 6. This will identify your most prominent behavioral features.

It would be likely that a person would not demonstrate some of these behaviors very often. Rate some of the items 1 or 2 to identify your least prominent behavioral characteristics.

Respond to the remaining items with 3 or 4 ratings depending upon how often or how characteristic you think these behaviors are of you.

Almost Never ① Not at All Characteristic	Seldom or Once in a while ② Slightly	Regularly but Not often. ③ Moderately Characteristic	Fairly Often ④ Characteristic	Very Frequently ⑤ Very Characteristic	Almost Always ⑥ Extremely

1A. Shows a sincere interest in suggestions from members of the work group. ①②③④⑤⑥

APPENDIX B

Forced distribution instructions to respondents on the Employee Profile.

Read each item and determine how characteristic of you the behavior associated with the item is using the "1" to "10" scale. In order to ensure that your most and least characteristic traits are identified, rate at least twenty (20) items "1" or "2", and no more than twenty (20) items at "9 or 10." You may find that putting 5 items in the "1" or "2" and 5 in the "9" or "10" ranges on EACH page is easiest way to accomplish this. Rate the remaining items within the range "3" and "8" depending upon your assessment of how characteristic of you each item is.

Less Characteristic	①②	③④⑤⑥⑦⑧	⑨⑩	More Characteristic

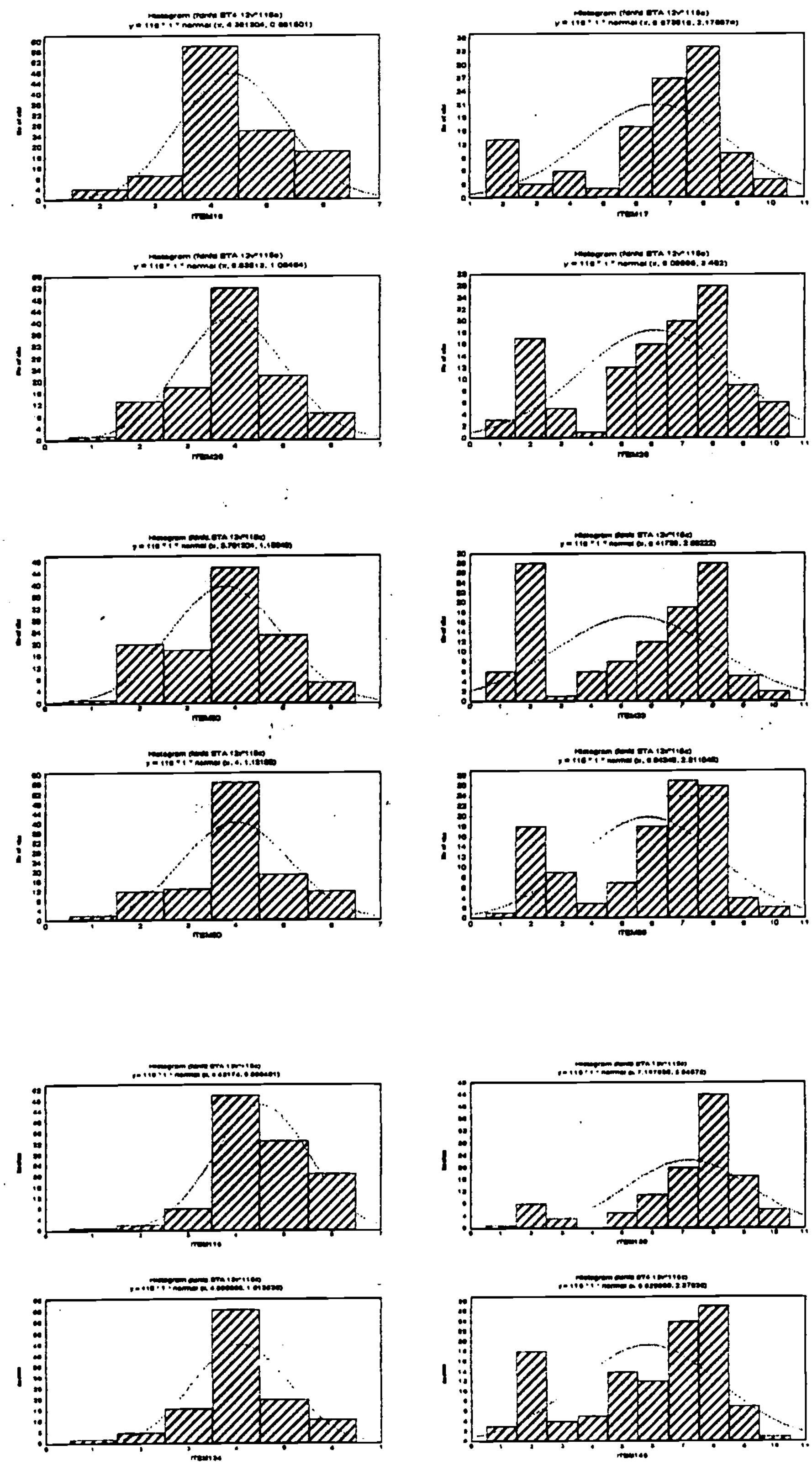
Shows a sincere interest in suggestions from members of the work group ①② ③④⑤⑥⑦⑧ ⑨⑩

BEST COPY AVAILABLE

APPENDIX C

Frequency histograms for items 1-6 (raw data)

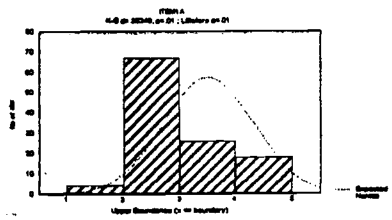
6-point scale | 10-point scale

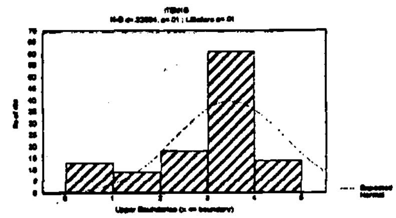


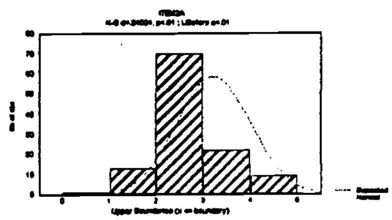
APPENDIX C, cont.

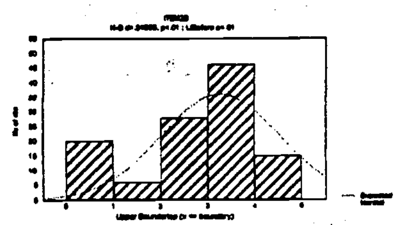
Frequency histograms for items 1-6 (collapsed data)

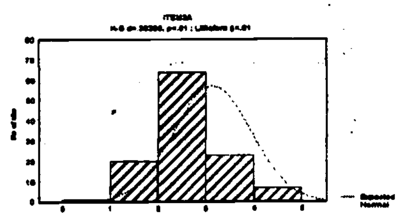
collapsed 6-point scale | collapsed 10-point scale

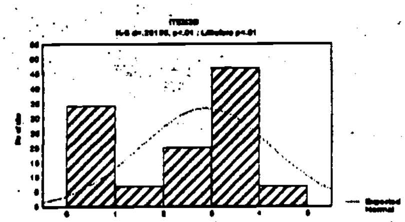


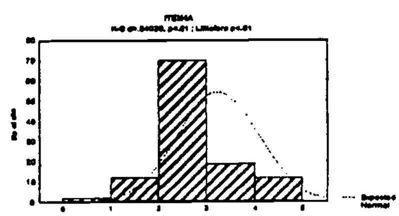


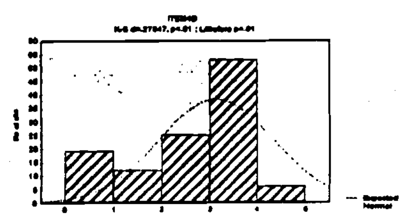


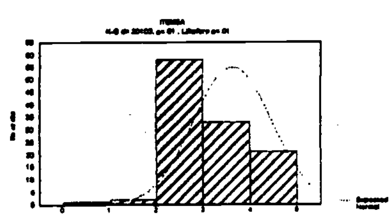


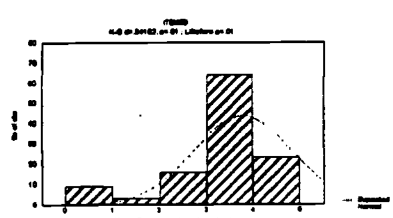


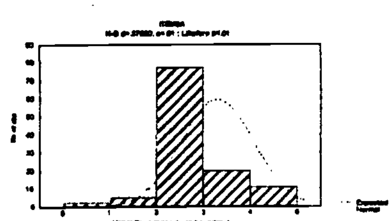


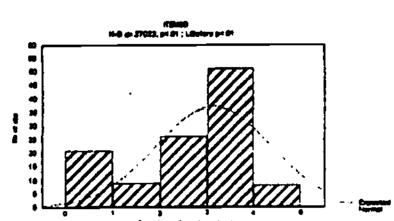








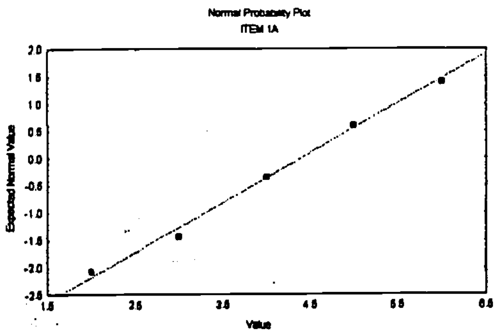




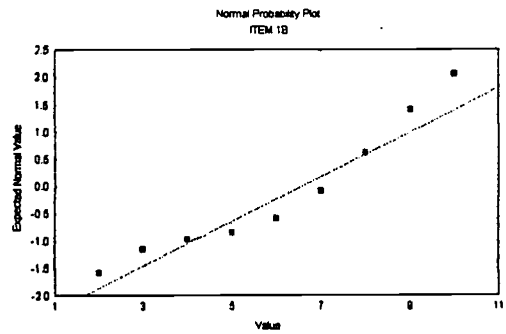
APPENDIX D

Normal probability plots for Item 1

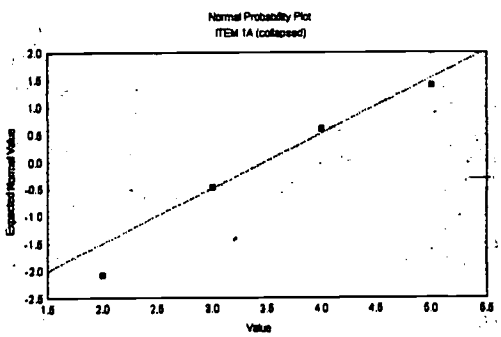
a. Raw data, NFD set.



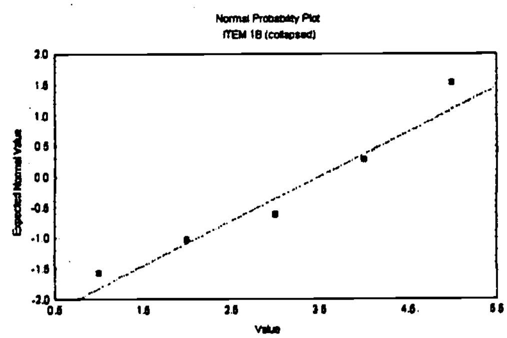
b. Raw data, FD set.



c. Collapsed data, NFD set.



d. Collapsed data, FD set.



APPENDIX E

Raw scale means and dispersion.

Non-forced distribution; 6-point scale (1-6)

	Mean	Standard Deviation	Standard Error
ITEM 1A	4.391	0.961	.090
ITEM 2A	3.939	1.095	.102
ITEM 3A	3.791	1.159	.108
ITEM 4A	4.000	1.132	.105
ITEM 5A	4.522	0.985	.092
ITEM 6A	4.087	1.014	.095

Forced distribution; 10-point scales (1-10)

	Mean	Standard Deviation	Standard Error
ITEM 1B	6.574	2.177	.203
ITEM 2B	6.096	2.492	.232
ITEM 3B	5.417	2.662	.248
ITEM 4B	5.843	2.312	.216
ITEM 5B	7.148	2.049	.191
ITEM 6B	5.826	2.370	.221

AERA

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM032241

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: THE IMPACT OF INSTRUCTIONAL SET ~~ON~~ OF SELF-REPORT RATINGS ON A SURVEY OF PERSONALITY CHARACTERISTICS

Author(s): JO-ANNE McDONALD & LISA HALL

Corporate Source: UNIV. OF DENVER

Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1	Level 2A	Level 2B
☑	☐	☐
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here,→ please

Signature: J McDonald	Printed Name/Position/Title: JO-ANNE McDONALD	
Organization/Address: UNIV. OF DENVER, DENVER, CO 80208	Telephone: 303-221-6787	FAX:
	E-Mail Address: jmcdonal@du.edu	Date: 6-3-2000

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
UNIVERSITY OF MARYLAND
1129 SHRIVER LAB
COLLEGE PARK, MD 20772
ATTN: ACQUISITIONS

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com